


Le distribuzioni statistiche

17 marzo 2008
prof. Fabio Bonoli

Sommario

1. Che cosa è la statistica?
2. La statistica descrittiva: riassumere le informazioni numeriche
3. La statistica descrittiva: interpolazione statistica, correlazione e regressione
4. Variabili casuali
5. Distribuzioni di probabilità

La statistica ha esigenze di tipo:



Descrittivo: corrisponde al bisogno di raccogliere dati e di fornire una sintesi

Predittivo: fornendo una fotografia del passato e del presente, aiuta a prevedere i comportamenti futuri.

La descrizione dei dati o "statistica descrittiva"

In misura crescente, le informazioni sull'ambiente che ci circonda sono disponibili in forma *quantitativa*.

I dati di cui possiamo disporre - i dati di un censimento, il reddito e il consumo pro-capite delle migliaia di intervistati in un'indagine - per la loro quantità non consentono una comprensione immediata di un fenomeno, e devono essere in qualche modo riassunti e sintetizzati.

La **statistica descrittiva** si interessa a questo problema.

Quel che si descrive è la popolazione.

La **popolazione** è l'insieme di tutte le misurazioni che ci interessano.

Il campionamento e l'inferenza statistica

Molto spesso, per analizzare un problema non è possibile considerare l'intera popolazione rilevante, perché essa è troppo numerosa rispetto alle risorse a disposizione per l'indagine.

Ci si deve allora accontentare di prendere in considerazione un sottoinsieme della popolazione detto **campione**.

Questo è il tema affrontato **dall'inferenza statistica**, che si occupa dell'utilizzo di un sottoinsieme della popolazione con l'obiettivo di trarre delle conclusioni che riguardino una popolazione nel suo insieme, per fini di conoscenza o per migliorare l'efficacia delle nostre decisioni.

Esempio: Siamo interessati alla statura media degli abitanti di Cesena. Consideriamo un ***campione*** formato da 100 abitanti di Cesena, estratti a sorte dalla ***popolazione*** dei suoi 94000 abitanti, calcoliamo la loro altezza media per avere un'idea dell'altezza media della popolazione.

L'estrazione di un campione dalla popolazione è chiamata **campionamento**, e deve essere realizzata in modo opportuno affinché si possa effettuare inferenza statistica in modo corretto. Le varie tecniche per estrarre un campione da una popolazione sono l'oggetto della **teoria dei campioni**, la più semplice di queste, è il **campionamento casuale semplice** che consiste nell'estrazione dalla popolazione.

Relazioni tra fenomeni e le previsioni

Nell'ambiente che ci circonda raramente i fenomeni esistono in isolamento, spesso essi sono collegati tra loro da relazioni di reciproca dipendenza.

Esempio: Le condizioni meteorologiche a Forlì verosimilmente sono "collegate" con le condizioni meteorologiche a Cesena, distante da essa solo 20 km.

Esempio: Il consumo di un individuo è "correlato" con il suo reddito, nel senso che in media maggiore è il reddito, maggiore è il consumo.

La statistica permette di mettere in relazione più variabili con il fine di studiare la loro relazione reciproca. Una volta che si è stabilita la natura di questa relazione - che può essere tra variabili distinte, ma anche per esempio tra una variabile e il suo passato - essa può essere utilizzata per effettuare delle *previsioni* sul futuro.

Esempio: Si conoscono i dati del consumo e del reddito di un paese. Utilizzando questi dati, è possibile calcolare una previsione futura del consumo, dato un certo valore futuro del reddito.

Tra le numerose tecniche statistiche che si occupano di questi problemi, noi dedicheremo la nostra attenzione alla cosiddetta **analisi di regressione**.

Relazioni di causalità tra diversi fenomeni

Un tema strettamente collegato all'analisi delle relazioni tra diversi fenomeni riguarda l'individuazione di eventuali relazioni di causalità tra essi, vale a dire, nell'identificare una o più cause e uno o più effetti.

Ci limitiamo ad osservare che in nessun modo la presenza di un collegamento, o "correlazione", tra due fenomeni, implica la presenza di un rapporto di causa ed effetto tra essi.

Esempio: Il fatto che al pranzo di Natale solitamente sia riunita tutta la famiglia e si mangi il panettone, non implica né che la presenza di un panettone richiama a raccolta tutti i famigliari, né che le riunioni di famiglia causano la presenza di un panettone.

Notiamo anche che neppure il fatto che un fenomeno tenda ad anticipare temporalmente un altro fenomeno implica necessariamente la presenza di una relazione di causalità dal primo verso il secondo.

Esempio: gli auguri di Natale anticipano il Natale, ma non lo causano.

Introduzione alla statistica descrittiva

Una volta che si sono raccolti i dati relativi a un certo fenomeno, questi vengono organizzati in vario modo e registrati su carta o, sempre più spesso, in formato elettronico.

Spesso i dati sono in grande quantità, al punto che non è possibile analizzarne le caratteristiche semplicemente leggendoli.

Esempio: Siamo interessati ai risultati delle ultime elezioni politiche in Italia e per questo ci vengono forniti i dati contenenti i risultati per ognuno degli oltre 8000 comuni, la lettura dei dati non ci permette di farci un'idea d'insieme del fenomeno che ci interessa.

La statistica descrittiva si pone l'obiettivo di descrivere e di riassumere certe caratteristiche rilevanti dei dati. Nel caso delle elezioni, per esempio, probabilmente quello che ci interessa non sono i dati comunali "grezzi", ma piuttosto le percentuali medie nazionali di ciascuna forza politica.

I dati possono essere di tipo **qualitativo o quantitativo**. Sono qualitativi i dati che si prestano a suddivisione per categorie, per esempio le risposte "sì" o "no". Sono quantitativi i dati espressi in forma numerica.

Introduzione alla statistica descrittiva

I dati **qualitativi** possono a loro volta essere **ordinabili**, se esiste un nesso logico per disporli, oppure **non ordinabili**, nel caso contrario.

Esempio: Le risposte alla domanda "Lei quanto lavora mediamente ?" sono: "per nulla, poco, ne molto ne poco, molto, moltissimo" costituiscono un insieme di dati qualitativi e ordinabili (da "per nulla" a "moltissimo").

Esempio: Le risposte alla domanda: "Dove preferisce trascorrere le sue vacanze?" sono: "al mare, ai monti, altrove" costituiscono un insieme di dati qualitativi non ordinabili: non è possibile individuare un ordine logico alle risposte.

I dati quantitativi esprimono una misurazione o un conteggio.

Si può distinguere tra dati **quantitativi continui** e **quantitativi discreti**.

Sono quantitativi continui le risposte numeriche che possono assumere qualsiasi valore all'interno di un intervallo, discreti quando variano mediante "salti".

Esempio: Le risposte alla domanda "Quante sigarette ha fumato ieri?" costituiscono dati quantitativi discreti, trattandosi di un conteggio di unità che possiamo considerare indivisibili.

Misure di tendenza centrale dei dati

Supponiamo di visitare un *pub* per 9 giorni consecutivi, e di registrare il numero di clienti entrati dalle ore 22 sino a mezzanotte. Per facilitarne la lettura, disponiamo le osservazioni in ordine crescente:

145, 160, 161, 200, 205, 210, 240, 270, 290

Il fenomeno varia di giorno in giorno, però è possibile individuare quel che potremmo definire un "centro" dei dati: uno o più valori che tendono a situarsi al centro - in qualche senso - della variazione del numero degli avventori nel locale, e che quindi tendono ad essere lontani da entrambi gli estremi, dunque in un certo senso ne troppo alti ne troppo bassi.

Esistono diverse misure che cercano di descrivere la "tendenza centrale" dei dati. Queste misure sono molto utili perché la "tendenza centrale" è di per sé una caratteristica interessante e rilevante di un certo fenomeno.

Per esempio, il gestore del *pub* che deve ordinare con un certo anticipo i fusti di birra per la sua clientela, è interessato magari non tanto al numero di clienti di giorno in giorno, quanto al loro numero tendenziale o "medio".

La **media aritmetica** è il principale indicatore della tendenza centrale dei dati.

Le medie

Definizione generale:

Una media è un numero che sintetizza una distribuzione statistica semplice

Il modo per sintetizzare una distribuzione non è unico .

Definizione di media:

Data una distribuzione x_1, x_2, \dots, x_n ; una sua media è la quantità

\bar{x} che, se sostituita a ciascun termine della distribuzione è tale che:

$$f(x_1, x_2, \dots, x_n) = f(\bar{x}, \bar{x}, \dots, \bar{x})$$

Esempi:

1. f è la somma dei termini

$$x_1 + x_2 + \dots + x_n = \bar{x} + \bar{x} + \dots + \bar{x} = n \cdot \bar{x} \text{ da cui}$$

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \text{ ovvero la media aritmetica}$$

Le medie

2. f è il prodotto dei termini

$$x_1 \cdot x_2 \cdot \dots \cdot x_n = \bar{x} \cdot \bar{x} \cdot \dots \cdot \bar{x} = \bar{x}^n \quad \text{da cui}$$

$$\bar{x} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} \quad \text{ovvero la media geometrica}$$

3. f è la somma dei quadrati dei termini

$$x_1^2 + x_2^2 + \dots + x_n^2 = \bar{x}^2 + \bar{x}^2 + \dots + \bar{x}^2 = \bar{x}^2 \cdot n \quad \text{da cui}$$

$$\bar{x} = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}} \quad \text{ovvero la media quadratica}$$

4. f è la somma dei reciproci dei termini

$$\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} = \frac{1}{\bar{x}} + \frac{1}{\bar{x}} + \dots + \frac{1}{\bar{x}} = \frac{n}{\bar{x}} \quad \text{da cui}$$

$$\frac{1}{\bar{x}} = \frac{\sum_{i=1}^n \frac{1}{x_i}}{n} \quad \text{e quindi} \quad \bar{x} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad \text{ovvero la media armonica}$$

Le medie

Oltre alle medie analitiche vi sono le medie lasche, quelle che si determinano

- basandosi sulla posizione di uno o alcuni termini
- facendo dei calcoli solo su alcuni termini della distribuzione.

Considerando la distribuzione ordinata in modo crescente, le più importanti medie lasche sono:

1. *Valore centrale*

$$V_C = \frac{x_1 + x_n}{2}$$

2. *Moda* *modalità che presenta la massima frequenza*

3. *Mediana* *il termine che bipartisce la graduatoria in modo da lasciare alla sua sinistra lo stesso numero di termini che lascia alla sua destra.*

Proprietà delle medie

Le due proprietà sotto indicate valgono per tutti i tipi di media

- Considerando una graduatoria la media è interna ad essa $x_1 \leq \bar{x} \leq x_n$
- Ogni media è espressa nella stessa unità di misura in cui sono espressi i termini della distribuzione.

Proprietà della media aritmetica

- La somma algebrica degli scarti dei termini della distribuzione dalla media aritmetica è 0

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

- La somma dei quadrati degli scarti dei termini della distribuzione dalla media aritmetica è minima

$$\sum_{i=1}^n (x_i - \bar{x})^2 \text{ minimo}$$

Proprietà delle medie

Proprietà della mediana

La somma dei valori assoluti degli scarti dei termini della distribuzione dalla mediana è minima

$$\sum_{i=1}^n |x_i - Me| \text{ minimo}$$

Osservazioni

- Una distribuzione ha media oggettiva se è costituita da misure ripetute di una stessa grandezza; in questo caso la media assume il significato di stima. Negli altri casi si parla di media soggettiva. Se una distribuzione ha media oggettiva si usa la media aritmetica per stimare la “vera misura”
- Rispetto alla media aritmetica, la media geometrica dà meno peso ai valori anomali.

Misure di variabilità dei dati

Consideriamo nuovamente gli avventori del *pub* durante il periodo di 9 giorni consecutivi:

145, 160, 161, 200, 205, 210, 240, 270, 290 per i quali avevamo la media è uguale a 209 avventori al giorno.

Consideriamo ora un altro ipotetico *pub*, per il quale, nei medesimi giorni, siano stati contati i seguenti avventori, in ordine crescente:

206, 206, 207, 208, 209, 210, 211, 212, 212

Le due distribuzioni hanno la stessa media: 209.

Se volessimo descrivere l'afflusso dei clienti nei due *pub* solo attraverso la tendenza centrale concluderemmo che i due fenomeni sono uguali.

Le due distribuzioni sono caratterizzati da una diversa **dispersione**: relativamente ampia nel primo caso, ridotta nel secondo.

La variabilità può definirsi come l'attitudine di un carattere ad assumere diverse modalità (varrà 0 se i termini sono tutti uguali, crescerà al crescere della diversità tra i termini)

Campo di variazione e varianza

Una misura di dispersione assai semplice è il **campo di variazione**, definito come la differenza tra l'osservazione più grande e l'osservazione più piccola:

Il campo di variazione è un indice molto semplice, ma fornisce una prima indicazione riguardo all'ammontare della dispersione di un insieme di dati.

Il campo di variazione risolve questo problema in modo assai semplice, trascura tutte le osservazioni escluso la maggiore e la minore

Possiamo immaginare delle misure di dispersione che considerino in modo esplicito tutte le osservazioni a disposizione, e le loro distanze da una misura di tendenza centrale, per esempio la media aritmetica. Una siffatta misura di dispersione è la **varianza**, di gran lunga la misura di dispersione più nota e usata.

Un'altra misura della dispersione è lo **scarto quadratico medio**, uguale alla radice quadrata della varianza.

$$CV = \frac{x_{\max} - x_{\min}}{2}$$

$$Var(X) = \sigma^2 = \frac{\sum (x_i - m)^2}{n}$$

Interpolazione statistica

Si parla di interpolazione quando:

Siano x e y due distribuzioni statistiche, le coppie di dati $(x; y)$ sono interpretabili come punti di un piano; ci si propone di costruire una funzione, detta **funzione interpolante**, che sia in grado di descrivere la relazione che intercorre fra l'insieme dei valori di x e l'insieme dei valori di y .

La ricerca della funzione interpolante può essere vista in due modi diversi.

Ci troviamo in presenza di una distribuzione che riteniamo lacunosa, cioè mancante di alcuni dati che non possono essere più rilevati. In questo caso la funzione interpolante deve servire a stimare i dati mancanti. **OPPURE**

Ci troviamo in presenza di una distribuzione di dati alcuni dei quali vengono ritenuti affetti da errori. In questo caso la funzione interpolante deve servire a sostituire ai dati che si ritengono affetti da errori dati che si ritengono più attendibili. In sostanza, alla distribuzione di partenza viene sostituita una distribuzione approssimata ma non affetta da errori.

Naturalmente, in entrambi i casi occorre, anzitutto, decidere il tipo di funzione che si vuole adottare: funzione lineare, quadratica, esponenziale, ecc.

Interpolazione statistica: metodo dei minimi quadrati

Consideriamo un fenomeno statistico per il quale si dispone della seguente distribuzione di dati:

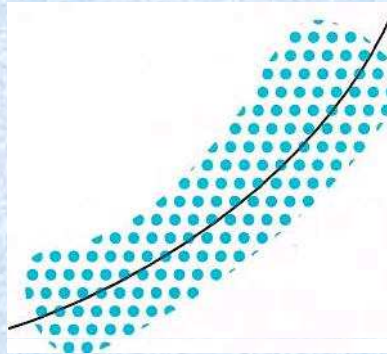
X	y
x_1	y_1
x_2	y_2
...	...
x_n	y_n

Poiché, come già detto, riteniamo che i dati osservati sono affetti da errori vogliamo costruire una funzione che ci permetta di sostituire ai dati y_i , osservati i dati y'_i interpolati, più regolari anche se approssimati. In questo caso, il problema che si presenta è quello di scegliere la funzione interpolante in modo che l'approssimazione presenti un «buon accostamento».

L'adozione del tipo di funzione può essere suggerita dall'esame del diagramma a dispersione.

Interpolazione statistica: metodo dei minimi quadrati

Se il diagramma a dispersione si presenta come in figura, sembra più logica l'adozione di una funzione esponenziale, anche se viene spesso adottata una funzione lineare a causa della sua semplicità.



Condizione per un buon accostamento

Ci chiediamo come bisogna procedere affinché, una volta scelto il tipo di funzione interpolante, possa ottenersi un buon accostamento fra la distribuzione dei valori osservati e quella dei valori teorici ottenuti interpolando?

Poiché gli errori commessi sono dati dalle differenze:

$$y_i - y'_i$$

al fine di ottenere un buon accostamento occorre minimizzare questi errori.

Interpolazione statistica: metodo dei minimi quadrati

Naturalmente, per raggiungere l'obiettivo voluto non possiamo prendere in considerazione la somma delle differenze: essendo alcune differenze positive e altre negative, esse potrebbero anche compensarsi. Consideriamo la somma dei quadrati delle differenze e poniamo come condizione di accostamento la seguente:

$$\sum_{i=1}^n (y_i - y'_i)^2 \text{ minimo}$$

Quindi, si parla di interpolazione statistica col metodo dei minimi quadrati quando, qualunque sia il tipo di funzione interpolante adottato, la condizione di accostamento fra valori osservati e valori teorici rende minima la somma dei quadrati delle differenze fra valori osservati e valori teorici.

Se come funzione interpolante viene scelta la funzione lineare:

$y = a + b \cdot x$ la condizione di accostamento è

$\sum (y_i - (a + bx_i))^2$ deve essere minima

Pertanto si deve minimizzare una funzione a 2 variabili $f(a, b)$

Si ricava $b = \frac{\sum (x_i - M_x)(y_i - M_y)}{\sum (x_i - M_x)^2}$; $a = M_y - bM_x$

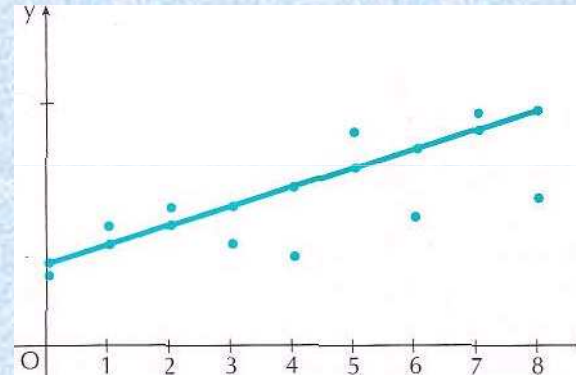
Interpolazione statistica: metodo dei minimi quadrati

Consideriamo i dati della tabella: produzione di frumento (in milioni di quintali) per gli anni dal 1987 al 1995.

X		
1987	0	85.65
1988	1	91.80
1989	2	94.60
1990	3	88.55
1991	4	86.40
1992	5	98.20
1993	6	90.50
1994	7	99.35
1995	8	90.70

Allora la retta interpolante è la seguente:

$$y = 0.738x + 88.798$$



Si può semplificare il calcolo di a, b

$$b = \frac{n \sum (x_i \cdot y_i) - \sum x_i \cdot \sum y_i}{n \sum (x_i)^2 - (\sum x_i)^2}; \quad a = \frac{\sum y_i \cdot \sum (x_i)^2 - \sum (x_i) \sum (x_i \cdot y_i)}{n \sum (x_i)^2 - (\sum x_i)^2}$$

Correlazione e regressione

Consideriamo due distribuzioni indicate con X e Y .

Esempio: $X =$ reddito mensile $Y =$ spesa mensile per vitto

E' naturale pensare che la spesa Y per vitto dipenda dal reddito X .

Quale relazione che lega Y a X ?

Primo modo. Riportiamo sull'asse delle ascisse i valori x_i e su quello delle ordinate i valori y_i . In questo modo otteniamo il diagramma della figura 1, detto *diagramma scatter*. Dal suo esame emerge l'esistenza della seguente relazione: a valori più alti di X corrispondono, in genere, valori più alti di Y .

Secondo modo. Riportiamo sull'asse delle ascisse i valori degli scarti $x_i - M_x$ e su quello delle ordinate i valori degli scarti $y_i - M_y$. In questo modo otteniamo il diagramma scatter di cui alla figura 2, Dal suo esame emerge l'esistenza della seguente relazione: salvo un caso, a scarti positivi di X (valori maggiori della media) corrispondono scarti positivi di Y e a scarti negativi di X (valori minori della media) corrispondono scarti negativi di Y .

In entrambi i casi il tipo di relazione che emerge è, ovviamente, lo stesso: **alla tendenza a crescere di uno dei due fenomeni corrisponde la tendenza a crescere anche dell'altro.**

Correlazione e regressione

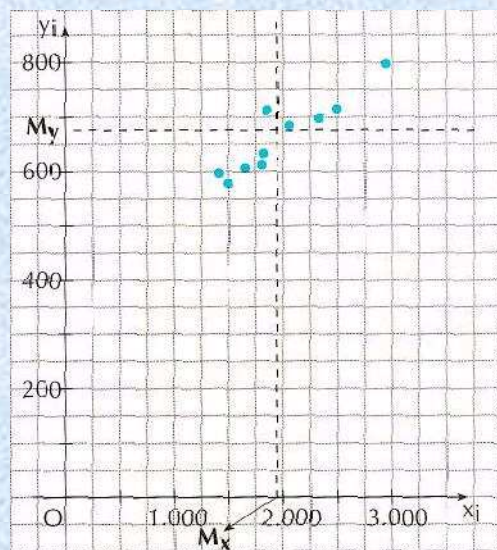


Fig. 1

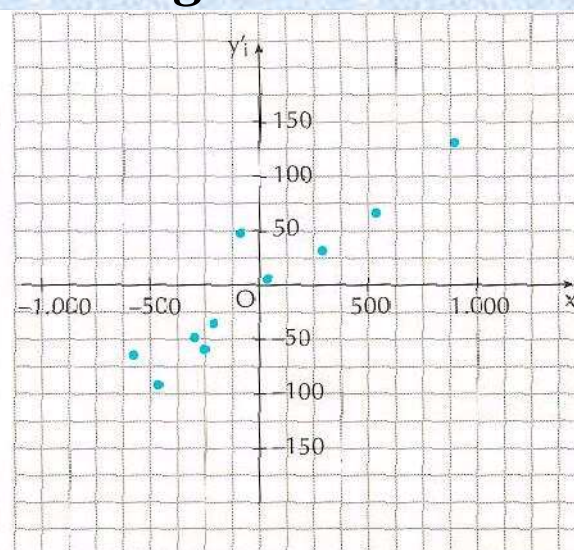


Fig 2

Fra X e Y esiste una relazione diretta se a valori crescenti di X corrispondono, prevalentemente, valori crescenti di Y mentre esiste una relazione inversa se a valori crescenti di X corrispondono, prevalentemente, valori decrescenti di Y.

Analogamente, ragionando in termini di scarti, diciamo che:

- fra X e Y esiste una relazione diretta se a scarti positivi (negativi) di X corrispondono, prevalentemente, scarti positivi (negativi) di Y. In questo caso i punti che rappresentano le coppie $(x_i; y_i)$ di scarti si addensano nel I e III quadrante;

Correlazione e regressione

Definizione: Si dice covarianza fra X e Y la media aritmetica dei prodotti degli scarti (corrispondenti).

Pertanto, indicando la covarianza col simbolo $Cov(X, Y)$, si scrive:

$$Cov(X, Y) = \frac{\sum (x_i - M_x)(y_i - M_y)}{n}$$

Precisiamo ora che:

- se a scarti positivi (negativi) di X corrispondono scarti positivi (negativi) di Y la relazione lineare fra i due fenomeni è diretta. In questo caso la somma dei prodotti degli scarti è positiva e, quindi, si ha: $Cov(X, Y) > 0$
- se a scarti positivi (negativi) di X corrispondono scarti negativi (positivi) di Y la relazione lineare fra i due fenomeni è inversa. In questo caso la somma dei prodotti degli scostamenti è negativa e, quindi, si ha: $Cov(X, Y) < 0$

In definitiva, possiamo dire che: **valori positivi della covarianza indicano l'esistenza di una relazione lineare diretta mentre valori negativi della covarianza indicano l'esistenza di una relazione lineare inversa.**

Coefficiente di correlazione lineare

Covarianza uguale a zero vuol dire che fra i due fenomeni non esiste relazione di tipo lineare. Ciò, però, non esclude che possa esistere una relazione di altro tipo (per esempio, di tipo parabolico).

La covarianza presenta l'inconveniente di assumere valori senza limitazione alcuna. Questo fatto la rende poco idonea a dare un'idea precisa sulla relazione fra X e Y .

Per superare l'inconveniente si preferisce, allora, ricorrere a un indice diverso ottenuto dividendo la covarianza per il prodotto degli scarti quadratici medi σ_x e σ_y .

Si ottiene così il coefficiente di correlazione lineare r

$$r = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\sum (x_i - M_x)(y_i - M_y)}{n \cdot \sigma_x \cdot \sigma_y}$$

Il vantaggio offerto dal coefficiente di correlazione lineare è che risulta: $-1 \leq r \leq 1$

- se $r = 0$ non esiste alcuna relazione lineare fra X e Y (non è, però, escluso che possa esistere una relazione di altro tipo, per esempio parabolica).
- se $r = -1$ si ha perfetta correlazione lineare inversa o perfetta correlazione lineare negativa;
- se $r = +1$ si ha perfetta correlazione lineare diretta o perfetta correlazione lineare positiva.

Regressione

Abbiamo detto che la correlazione si occupa della misura del grado di intensità della relazione lineare che intercorre fra X e Y .

Per raggiungere questo obiettivo i due fenomeni vengono indagati in modo associato e simmetrico determinando il coefficiente di correlazione lineare: non si considera il fatto che Y dipenda da X o che X dipenda da Y .

Spesso, però, ciò che interessa sta proprio nell'assegnare a uno dei due fenomeni il ruolo di variabile indipendente e all'altro il ruolo di variabile dipendente nel qual caso si vuole vedere come varia il fenomeno dipendente rispetto al fenomeno indipendente.

In questo caso si parla di indagine asimmetrica e bisogna fare ricorso allo studio della regressione.

Esempio: Una ditta vende un dato prodotto. La quantità venduta varia al variare del prezzo come mostrano i seguenti dati:

prezzo	800	820	840	850	855
quantità	1.200	1.170	1.130	1.080	1.020

Regressione

La quantità venduta dipende dal prezzo e interessa sapere come varia la quantità al variare del prezzo.

D'altra parte, il coefficiente di correlazione lineare è $r = -0,932$ cioè una forte correlazione negativa fra quantità Y venduta e prezzo X , non permette di dare risposta a un quesito del tipo:

qual è la quantità venduta se il prezzo è uguale a 832? Per dare una risposta a questa domanda occorre necessariamente fare ricorso allo studio della regressione di Y rispetto a X .

Regressione lineare col metodo dei minimi quadrati

Consideriamo, in generale, due fenomeni X e Y per i quali possiamo stabilire che:

X è la variabile indipendente o esplicativa Y è la variabile dipendente

Il problema è sostanzialmente analogo a quello già visto per l'interpolazione statistica:

$$y = a + b \cdot x$$

$$\text{Si ricava } b = \frac{\sum (x_i - M_x)(y_i - M_y)}{\sum (x_i - M_x)^2} = r \cdot \frac{\sigma_y}{\sigma_x}; \quad a = M_y - bM_x$$

Variabili casuali

Innanzitutto, consideriamo il concetto di *variabile casuale*

Una **variabile casuale** è una risposta, dipendente dal caso, che otteniamo a una determinata domanda :

Variabili casuali qualitative, esprimibili in forma categoriale:

Esempio: È qualitativa la risposta a una domanda di un questionario a cui si risponde con "Si" o con "No".

Variabili casuali quantitative, esprimibile in forma numerica. Queste possono essere distinte in:

discrete, se rappresentano il risultato di una numerazione:

Esempio: Il numero di "teste" su 4 lanci di una moneta

oppure **continue**, negli altri casi:

Esempio: La statura di un individuo, il debito pubblico italiano l'anno prossimo.

Variabili casuali discrete semplici

Una **variabile casuale discreta semplice** è una variabile X che, in un esperimento a k eventi necessari e incompatibili A_1, A_2, \dots, A_k assume x_1, x_2, \dots, x_k

Esempio: Sia la prova o esperimento: lancio di 3 monete. Gli eventi necessari e incompatibili sono $A_1=TTT$ $A_2=TTC$... , $A_8=CCC$. La variabile casuale discreta X è una variabile che può assumere valori x_i (ad esempio il numero di teste)

X	$x_1 = 0$	$x_2 = 1$	$x_3 = 2$	$x_4 = 3$
P	1/8	3/8	3/8	1/8

Ai valori assunti da una variabile casuale sono associate probabilità.

Si dice funzione di ripartizione per una variabile casuale discreta:

$$F_i = F(x_i) = P(X \leq x_i) = \sum_{k=1}^i p_k$$

Una funzione con valori tra 0 e 1, non decrescente e tale che:

$$P(a \leq X \leq b) = F(b) - F(a)$$

Valore atteso o speranza matematica

Il valore atteso di una variabile casuale quantitativa discreta, o speranza matematica è la media aritmetica di una distribuzione di probabilità di una variabile casuale x , ed è indicata con $E(x)$. È dato dalla somma dei prodotti tra il valore della variabile casuale e la probabilità ad essa associata:

$$E(x) = \sum_{i=1}^n x_i \cdot p(x_i)$$

dove n è il numero di modalità che x , può assumere.

Esempio: Il valore atteso del risultato del lancio di un dado è uguale a:

$$E(x) = 1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 21/6 = 3.5$$

Esempio: Consideriamo una scommessa in cui con probabilità $1/10$ si vincono 7 € e nei rimanenti casi si perde 1 €.

Il valore atteso della scommessa è $7 \text{ €} \cdot 1/10 + (-1 \text{ €}) \cdot 9/10 = -0.2 \text{ €}$.

Varianza di una variabile casuale discreta

La **varianza** di una variabile casuale esprime la sua dispersione attorno al suo valore atteso:

$$Var(x) = \sigma^2(x) = \sum_{i=1}^n (x_i - M_x)^2 \cdot p(x_i)$$

dove n è il numero di modalità che x ; può assumere.

Esempio: la varianza del risultato del lancio di un dado è uguale a:

$$Var(X) = (1-3.5)^2 \cdot 1/6 + (2-3.5)^2 \cdot 1/6 + (3-3.5)^2 \cdot 1/6 + (4-3.5)^2 \cdot 1/6 + (5-3.5)^2 \cdot 1/6 + (6-3.5)^2 \cdot 1/6 = 17.5$$

Lo **scarto quadratico medio**, indicato con σ è la radice quadrata della varianza.

Esempio: lo scarto quadratico medio del risultato del lancio di un dado è 4.18.

Una coppia di variabili casuali semplici (X, Y) definisce una **variabile casuale doppia**. La sua rappresentazione è una tabella a doppia entrata dove nella cella di posto i,k vi è la probabilità p_{ik}

In particolare tra X e Y vi è indipendenza se $p_{ik} = p_i \cdot p_k$

Variabili casuali continue

Variabile casuale continua è una variabile casuale che assume tutti i valori reali compresi in un intervallo limitato o illimitato

Esempio: consideriamo una persona che attualmente ha 30 anni, la *durata di vita residua* è una variabile casuale continua, egli può morire in un qualsiasi istante facente parte dell'intervallo $x > 30$

La **funzione di ripartizione** $F(x)$ è definita per ogni x reale, $F(a)=0$ e $F(b)=1$, è monotona non decrescente.

$F(x)$ = probabilità che la variabile casuale assuma un valore minore o uguale a x .

Se $F(x)$ è una funzione derivabile allora $f(x) = F'(x)$ prende il nome di **densità di probabilità**.

Alcune relazioni importanti: $F(x) = \int_a^x f(t)dt$ $\int_a^b f(t)dt = 1$

Essendo $F(x_1) = \int_a^{x_1} f(t)dt$ $F(x_2) = \int_a^{x_2} f(t)dt$

$\Pr(x_1 \leq x \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(t)dt$

Parte IV: Variabili casuali

Valore medio, varianza e scarto quadratico medio

Variabili casuali continue, alcune formule importanti:

$$M(X) = \int_a^b x \cdot f(x) dx \quad \text{valore medio}$$

$$\text{Var}(X) = M(X^2) - (M(X))^2$$

$$\sigma(X) = \sqrt{\text{Var}(X)}$$

La moda di una variabile casuale continua

è quel valore al quale corrisponde la massima densità di probabilità

La mediana di una variabile casuale continua si ottiene risolvendo $F(X) = 1/2$

Distribuzioni di probabilità

Una **distribuzione di probabilità** è un elenco di risultati tra loro reciprocamente incompatibili ed esaustivi dello spazio campionario e delle probabilità ad essi associate. La somma di queste probabilità è uguale a 1.

Esempio: Consideriamo gli esiti possibili del lancio di un dado non truccato. La distribuzione di probabilità è:

Risultato: 1 2 3 4 5 6 Probabilità: $1/6$ $1/6$ $1/6$ $1/6$ $1/6$ $1/6$

Si noti che le probabilità sommano a 1.

Esempio: Supponiamo che la probabilità che piova sia uguale a 0.2, e la probabilità che vi sia il sole sia uguale a 0.5. Le probabilità associate a questi eventi non costituiscono una distribuzione di probabilità.

Infatti, i due eventi non sono mutuamente esclusivi (talvolta piove e contemporaneamente vi è il sole), non sono esaustivi dello spazio campionario (vi sono condizioni meteorologiche oltre alla pioggia e al sole), e quindi le probabilità non sommano a uno.

La distribuzione binomiale

Consideriamo 4 lanci successivi di una moneta. Siamo interessati al numero di volte che, in 4 lanci, esce "Testa". Indichiamo dunque con "X" la variabile casuale discreta "numero di teste in 4 lanci di una moneta".

Vogliamo calcolare la probabilità associata a ciascuno di questi possibili valori di X (0, 1, 2, 3, 4).

La distribuzione di probabilità della variabile

$$X = 0 \quad 1 \quad 2 \quad 3 \quad 4$$

$$\text{Probabilità} = 0.0625 \quad 0.25 \quad 0.375 \quad 0.25 \quad 0.625$$

È immediato verificare che le probabilità sommano a 1.

E' un esempio di una variabile casuale distribuita secondo la distribuzione binomiale.

La distribuzione binomiale descrive le variabili casuali quantitative discrete che esprimono un numero di "successi", nell'esempio, il numero di "teste" nel lancio ripetuto di una moneta.

Perché il fenomeno sia descrivibile per mezzo della distribuzione binomiale, la probabilità di successo deve rimanere costante nel corso dell'esperimento.

Valore atteso e varianza nella distribuzione binomiale

Si può poi mostrare che il **valore atteso** di una variabile casuale distribuita secondo la distribuzione binomiale è dato da $E(x)=n \cdot p$; la **varianza** è data da $n \cdot p \cdot (1-p)$

Esempio: Al Bar Sport, la probabilità che, prima delle 9 di mattina, un cliente chieda un cappuccino, è uguale al 40%. Alle 8:30 nel bar sono presenti 8 clienti. Il valore atteso del numero di cappuccini richiesti dagli 8 clienti è dato da $E(X) = n p = 8 \cdot 0.40 = 3.2$; la varianza è data da $n p (1 - p) = 8 \cdot 0.40 \cdot (1 - 0.40) = 1.92$.

Supponiamo di voler conoscere la probabilità che, tra le 8 persone, 2 chiedano un cappuccino. Utilizziamo la formula della distribuzione binomiale:

$$P(X = k) = \binom{n}{k} (p)^k (1-p)^{n-k} \rightarrow P(X = 2) = \binom{8}{2} (0.4)^2 (0.6)^6 = 0.209$$

La probabilità invece che, per esempio, *almeno* 3 clienti chiedano un cappuccino è data da $P(X=3)+P(X=4)+P(X=5)+P(X=6)+P(X=7)+P(X=8)=1 - [P(X=0) + P(X=1) + P(X=2)]$

$$P(X = 0) = \binom{8}{0} (0.4)^0 (0.6)^8 = 0.017 \quad = 1 - 0.017 - 0.090 - 0.209 = 0.684$$

$$P(X = 1) = \binom{8}{1} (0.4)^1 (0.6)^7 = 0.090$$

Vi è una probabilità del 68.4% che su 8 clienti presenti nel bar almeno 3 chiedano un cappuccino.

La distribuzione ipergeometrica

Nell'esempio del lancio ripetuto di una moneta, l'esperimento era caratterizzato dal fatto che la probabilità di successo restava costante nel corso dell'esperimento

Supponiamo invece di effettuare da un'urna una *estrazione senza reinserimento*. Se effettuiamo *un'estrazione senza reinserimento* da un'urna con numerosità finita, la probabilità di successo non è costante, ma dipende dagli esiti precedenti.

Esempio: Consideriamo 9 persone in attesa dell'apertura di un bar. Di queste nove persone, 6 preferiscono un cappuccino e 3 un caffè.

La probabilità che una persona che entra preferisca un caffè non è costante, ma dipende dalle preferenze di chi lo ha preceduto dentro al bar. È uguale a 3/9 per il primo entrato, a 2/8 per il secondo, se il primo preferiva il caffè, oppure a 3/8 se invece preferiva il cappuccino, eccetera.

$$P(X = x | n, N, A) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}}$$

n è la dimensione del campione,
N la dimensione della popolazione,
x il numero di successi nel campione,
A il numero di successi nella popolazione.

La distribuzione ipergeometrica

Esempio: nel bar dell'esempio precedente, entrano 5 persone. Siamo interessati alla probabilità che 2 di questi preferiscano il caffè. In questo caso $N=9$, $n=5$, $A=3$, $x=2$.
Dunque,

$$P(X = 2 | n = 5, N = 9, A = 3) = \frac{\binom{3}{2} \binom{9-3}{5-2}}{\binom{9}{5}} = 0.476$$

Per la distribuzione ipergeometrica, la formula del valore atteso è uguale al caso della distribuzione binomiale: $E(X) = n p$; la **varianza** differisce dalla varianza del caso della distribuzione binomiale per un fattore di correzione:

$$s_x^2 = n p (1 - p) ((N - n) / (N - 1))$$

Se N è abbastanza grande, allora la probabilità di successo è circa costante e indipendente dalle estrazioni precedenti.

La distribuzione ipergeometrica può allora essere approssimata per mezzo della distribuzione binomiale, con un errore di approssimazione che sarà tanto più ridotto, quanto più grande è N .

La distribuzione normale

Consideriamo nuovamente l'esempio del lancio ripetuto di una moneta, dove la variabile casuale X , che rappresenta il numero di "teste" che si ottengono in un certo numero di lanci, è descritta dalla distribuzione binomiale.

Possiamo notare che al crescere di n la distribuzione tende ad assumere la forma di una *curva* con forma di *campana*, e che, a prescindere dal valore della probabilità di successo p , essa tende ad essere simmetrica attorno al proprio valore atteso.

Più grande è n , più la distribuzione della variabile casuale X tenderà ad essere una *curva continua* con forma di *campana* e simmetrica attorno al proprio valore atteso. Se n è molto grande, possiamo immaginare X come una variabile casuale che è il risultato della somma di n fattori diversi, ciascuno dei quali può essere presente con una probabilità, p .

È legittimo immaginare che molti fenomeni fisici o sociali possano essere immaginati come la *somma* di fattori molteplici, e che quindi la distribuzione che descrive X sia di particolare importanza.

Così è nei fatti: se immaginiamo che n possa essere infinitamente grande e che quindi X possa assumere un numero infinito di valori, allora essa non è più una variabile casuale *discreta*, ma *continua*. La distribuzione di X è allora la cosiddetta **distribuzione normale o gaussiana**.

La distribuzione normale

La distribuzione binomiale è una distribuzione di probabilità discreta, perché descrive le probabilità associate alle realizzazioni di una variabile casuale discreta. La distribuzione normale è invece una distribuzione continua.

Una **funzione di distribuzione di probabilità continua** è una funzione di distribuzione che descrive le probabilità associate a variabili casuali *continue*.

La probabilità che una variabile casuale continua assuma un valore compreso tra due intervalli a e b è uguale all'area al di sotto della sua distribuzione compresa tra quegli intervalli. L'area al di sotto di tutta la distribuzione è uguale a 1.

La **funzione di distribuzione normale** è caratterizzata da due parametri: il suo valore atteso μ_x e la sua varianza σ^2 .

La formula della funzione di distribuzione normale o, più propriamente, della "funzione di densità normale" è data da:

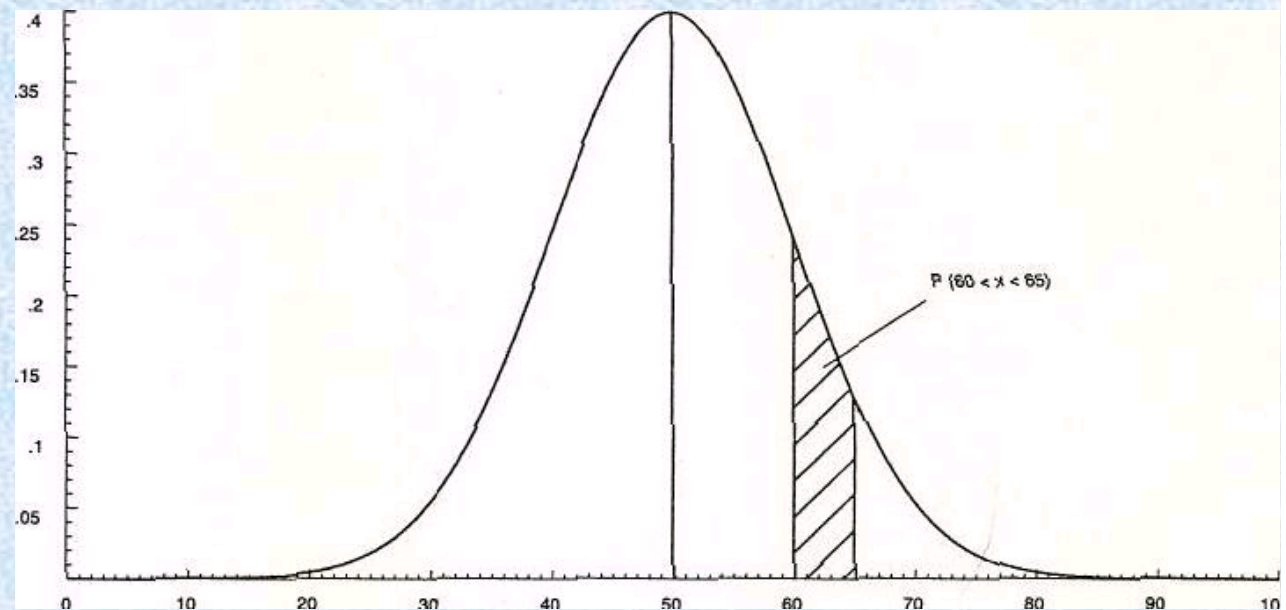
$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma_x}} e^{-\frac{1}{2} \frac{(x-\mu_x)^2}{\sigma_x^2}}$$

La distribuzione normale

Non è necessario conoscere, e neppure comprendere, la formula della distribuzione di probabilità normale per potere risolvere problemi che coinvolgano delle variabili casuali distribuite secondo la legge della distribuzione normale.

Una variabile casuale continua X distribuita in modo normale con valore atteso μ_x e la sua varianza σ^2 è indicata scrivendo $X \sim N(\mu_x, \sigma^2)$. La figura mostra la distribuzione di $X \sim N(\mu_x = 50, \sigma^2 = 100)$.

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} e^{-\frac{1}{2} \cdot \frac{(x-\mu_x)^2}{\sigma_x^2}}$$



Parte V: Distribuzioni di probabilità

Applicazioni della distribuzione normale

Si può pervenire alla variabile casuale gaussiana guardando come si distribuiscono gli errori riscontrabili nella misura ripetuta di una grandezza fisica.

Si riscontra che ciascuna misura è affetta da un errore.

Gli errori che si commettono nell'eseguire n misurazioni di una stessa grandezza sono determinati da numerose cause che non sono controllabili e che, quindi, non possono essere eliminate. Si tratta dei cosiddetti **errori accidentali**.

L'insieme delle cause non controllabili e, quindi, non eliminabili che determinano gli errori accidentali costituiscono «il caso». Proprio per questo motivo gli errori accidentali vengono detti anche **errori casuali**.

Diversi dagli errori casuali sono gli **errori sistematici** che, invece, sono dovuti a una causa ben precisa; per esempio, errori sistematici sono quelli causati dalla non perfetta taratura di uno strumento.

In questo caso, una volta individuata la causa che genera l'errore sistematico, si può fare una valutazione dell'errore che, pertanto, può essere rimosso.

E' possibile costruire un modello probabilistico in grado di descrivere il modo in cui gli errori casuali si distribuiscono?

La risposta è affermativa.

Gli errori casuali si distribuiscono secondo una variabile casuale gaussiana.

La distribuzione normale standardizzata

Per calcolare le probabilità associate a una variabile casuale continua distribuita normalmente con media μ_x e la sua varianza σ^2 è necessario calcolare l'area sottesa alla particolare distribuzione normale in corrispondenza di un dato intervallo.

Esempio, la probabilità che la variabile casuale $X \sim N(\mu_x = 50, \sigma^2 = 100)$ la cui distribuzione è mostrata in figura, assuma un valore compreso tra 60 e 65, è uguale all'area tratteggiata.

In linea di principio, il calcolo dell'integrale della funzione di probabilità normale dovrebbe essere realizzato ogni qualvolta si debba calcolare una probabilità relativa a una qualche variabile casuale distribuita normalmente.

Il problema del calcolo dell'integrale della funzione di distribuzione normale non è però di semplice soluzione.

Per ovviare a questo inconveniente, si è escogitato un modo che permette di ricondurre ogni particolare problema che coinvolge una variabile distribuita in modo normale a un problema particolare, detto della "variabile normale standardizzata", di cui si conosce la soluzione. E quindi sufficiente disporre di una regola, che permetta di trasformare un problema generico nel problema "standardizzato", per potere agevolmente risolvere qualsivoglia problema che coinvolga il calcolo di una probabilità di una variabile distribuita normalmente.

La distribuzione normale standardizzata

Consideriamo una variabile casuale $Z \sim N(\mu_x=0, \sigma^2=1)$ detta *variabile normale standardizzata*. Di questa variabile sono state calcolate un gran numero di aree, per intervalli che vanno da 0 a numerosissimi punti della retta.

I risultati di questo calcolo preventivo costituiscono le "tavole della funzione di distribuzione normale standardizzata"

La regola di *standardizzazione* consiste nel sottrarre da una variabile casuale X distribuita normalmente il suo valore atteso, dividere per il suo scarto quadratico medio, per ottenere la *variabile normale standardizzata*:

$$Z = \frac{X - \mu_x}{\sigma_x}$$

In questo modo, data una variabile X distribuita normalmente, possiamo facilmente esprimerla nei termini di un'altra variabile casuale, Z , essa pure distribuita normalmente, ma con valore atteso uguale a 0 e varianza unitaria.

L'utilizzo delle tabelle della distribuzione normale standardizzata

La tabella della distribuzione normale standardizzata fornisce l'area sottesa alla distribuzione di $Z \sim N(\mu_x=0, \sigma^2=1)$ in corrispondenza dell'intervallo da 0 a un determinato punto.

Parte V: Distribuzioni di probabilità

La distribuzione normale standardizzata

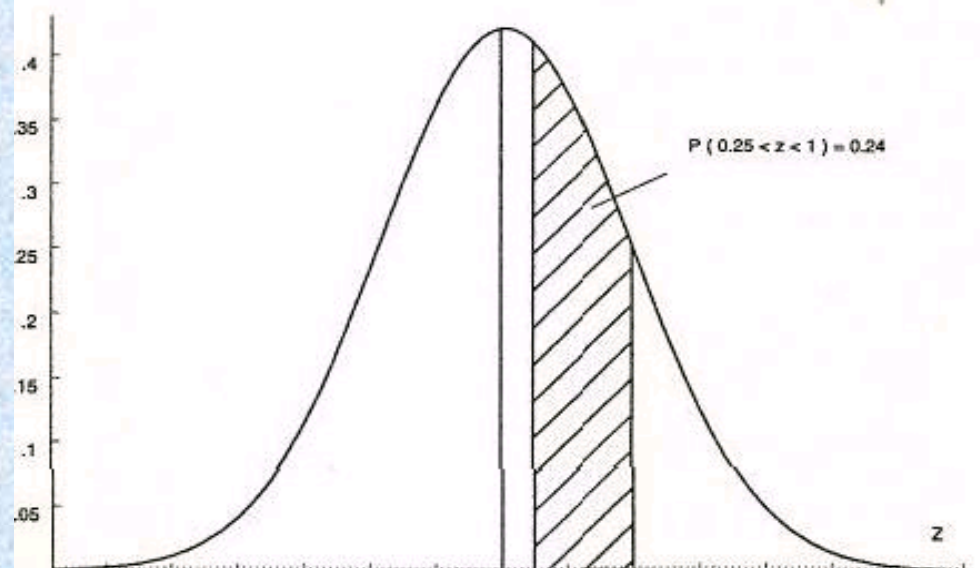
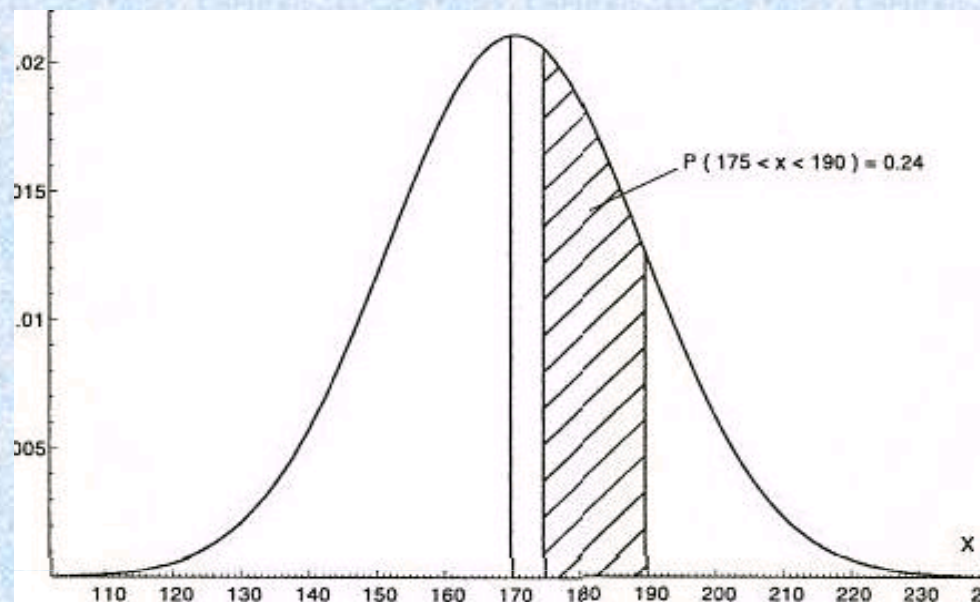
Esempio: La probabilità che la variabile casuale Z sia compresa tra 0 e 1.14 è uguale a 0.3729. La probabilità che la variabile casuale Z sia compresa tra 0 e -1.14 è 0.3729.

Se desideriamo calcolare la probabilità che Z sia compresa tra due punti entrambi diversi da 0, è necessario calcolare l'area tra 0 e il primo punto, quindi tra 0 e il secondo, e infine calcolare la differenza tra le due aree, se i due punti hanno lo stesso segno, oppure la loro somma, se i due punti hanno segno opposto.

Esempio: Calcoliamo la probabilità che Z sia compresa tra 1 e 1.14. La probabilità che Z sia compresa tra 0 e 1.14 è uguale a 0.3729; la probabilità che Z sia compresa tra 0 e 1 è uguale a 0.3423. Quindi la probabilità che Z sia compresa tra 1 e 1.14 è uguale a $0.3729 - 0.3423 = 0.0306$, ovvero circa il 3%.

Esempio: Calcoliamo la probabilità che Z sia compresa tra -1.23 e +1.14. La probabilità che Z sia compresa tra -1.23 e 0 è uguale a 0.3907; la probabilità che Z sia compresa tra 0 e 1.14 è uguale a 0.3729. Quindi, la probabilità che Z sia compresa tra -1.23 e 1.14 è uguale a $0.3907 + 0.3729 = 0.7636$.

La distribuzione normale standardizzata



Esempio: supponiamo che la statura degli abitanti di Cesena sia una variabile casuale X distribuita normalmente con media uguale a 170 centimetri, e varianza uguale a 400 centimetri: $X \sim N(\mu_x = 170, \sigma^2 = 400)$

Vogliamo calcolare la probabilità che un abitante di Cesena preso a caso abbia una statura compresa tra 175 cm. e 190 cm. La risposta è data dall'area al di sotto della funzione di distribuzione, rappresentata nella figura tra 175 e 190.

Non potendo agevolmente calcolare quell'area, procediamo con la *standardizzazione* degli estremi dell'intervallo, per trovare a che cosa corrispondono in termini della distribuzione normale standardizzata.

La distribuzione di Poisson

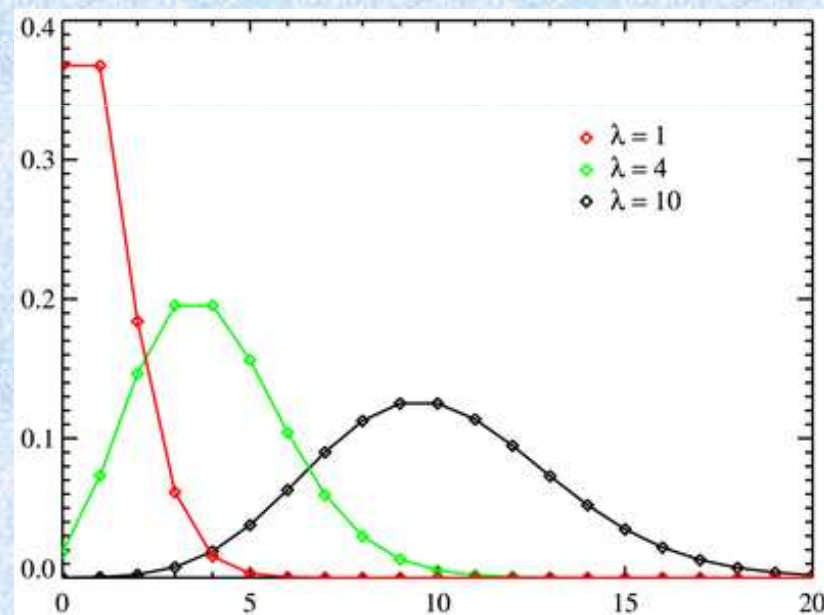
La *distribuzione di Poisson* è una distribuzione di probabilità che si presta a descrivere il numero di eventi relativamente poco frequenti all'interno di un certo periodo.

Per esempio, il numero di chiamate telefoniche in un ora, oppure, il numero di soldati Prussiani che ricevono un calcio da un asino in forza all'esercito nel periodo di un anno.

Sia la probabilità che assuma un generico valore k

$$p_k = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

$$E(X) = \lambda = \sigma^2$$



Bibliografia

Lucio Picci

Introduzione alla statistica CLUEB

Daniela Cocchi

Esercizi di statistica CLUEB

Trovato

Calcolo delle probabilità e statistica inferenziale G&C

www.istat.it

<http://www-eurisco.onecert.fr/>

<http://www.sis-statistica.it/>